Chapter 7

Conditional expectation

7.1 Inner product spaces

There are several ways to introduce the notion of conditional expectation. We begin by introducing inner-product spaces and motivate a definition of conditional expectation by using the Projection Theorem.

Definition 7.1. *A real vector space $\mathbb{X}$ is called an inner-product space if for all $x, y \in \mathbb{X}$, there exists a function $\langle x, y \rangle$, called an inner-product, such that for all $x, y, z \in \mathbb{X}$ and $a \in \mathbb{R}$*[1]

1. $\langle x, y \rangle = \langle y, x \rangle$
2. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
3. $\langle ax, y \rangle = a\langle x, y \rangle$, $a \in \mathbb{R}$
4. $\langle x, x \rangle \geq 0$, *for all* x
5. $\langle x, x \rangle = 0 \iff x = \theta$, *where* θ *is the null vector in* $\mathbb{X}$.

The following theorem shows that a general version of the Cauchy-Schwarz Inequality holds for inner-product spaces.

[1] If the vector space $\mathbb{X}$ is associated with a complex field, property 1 becomes $\langle x, y \rangle = \overline{\langle y, x \rangle}$, where for $x \in \mathbb{C}$, $\bar{x}$ is the complex conjugate of x, and in property 3 $a \in \mathbb{C}$.

Theorem 7.1. *Let $\mathbb{X}$ be an inner-product space and $x, y \in \mathbb{X}$. Then,*

$$|\langle x, y\rangle| \leq \langle x, x\rangle^{1/2}\langle y, y\rangle^{1/2}.$$

Proof. Let $y \neq \theta$ and note that for all $a \in \mathbb{R}$,

$$\begin{aligned} 0 &\leq \langle x - ay, x - ay\rangle = \langle x, x\rangle - 2a\langle x, y\rangle + a^2\langle y, y\rangle \\ &\leq \langle x, x\rangle - \frac{\langle x, y\rangle^2}{\langle y, y\rangle} \text{ by letting } a = \langle x, y\rangle/\langle y, y\rangle. \end{aligned}$$

The last inequality is equivalent to $\langle x, y\rangle^2 \leq \langle x, x\rangle\langle y, y\rangle$ or $|\langle x, y\rangle| = \langle x, x\rangle^{1/2}\langle y, y\rangle^{1/2}$. Lastly, if $y = \theta$ then the inequality holds with equality and $\langle x, \theta\rangle = 0$. ∎

It can be easily shown that the function $\|\cdot\| : \mathbb{X} \to [0, \infty)$ defined as $\|x\| = \langle x, x\rangle^{1/2}$ is a norm on $\mathbb{X}$. Thus, every inner-product space can be taken to be a normed space with this induced norm. Another important property in inner-product spaces is the Parallelogram Law, which is given in the next theorem.

Theorem 7.2. *In an inner-product space $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$.*

Proof. $\|x + y\|^2 = \langle x + y, x + y\rangle = \langle x, x\rangle + \langle y, y\rangle + 2\langle x, y\rangle$ and $\|x - y\|^2 = \langle x - y, x - y\rangle = \langle x, x\rangle + \langle y, y\rangle - 2\langle x, y\rangle$. Hence, we obtain

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

∎

Example 7.1. *Let $x, y \in \mathbb{R}^n$ and define $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$. It can be easily shown that $\langle x, y\rangle$ is an inner-product for $\mathbb{R}^n$ and $\langle x, x\rangle^{1/2} = \|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$ is a norm.*

Example 7.2. *Consider the space $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ of random variables $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ such that $\int_\Omega X^2 dP < \infty$. By Theorem 3.18.1 $XY \in \mathcal{L}(\Omega, \mathcal{F}, P)$ and by Theorem 3.18.3*

$\mathcal{L}^2(\Omega, \mathcal{F}, P)$ is a vector space. Now, define $\langle X, Y\rangle = E(XY) = \int_\Omega XY\,dP$. Using the properties of integrals, conditions 1-4 in Definition 7.1 are easily verified. However, condition 5 does not hold. Whereas it is true that $X(\omega) = 0$ for all ω, the null vector in $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, gives $\langle X, X\rangle = \int_\Omega X^2(\omega)dP = 0$, $\int_\Omega X^2(\omega)dP = 0$ does not imply $X(\omega) = 0$ for all ω. This is true since a random variable Z that takes non-zero values in sets of measure zero and is equal to 0 elsewhere will be such that $\int_\Omega Z^2(\omega)dP = 0$. If we treat any two variables X and Z in $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ as being identical if they differ only in a set of measure zero, that is if $P(\{\omega : X(\omega) \neq Z(\omega)\}) = 0$, then condition 5 is met and $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ is an inner product space with $\|X\|_2 = \left(\int_\Omega X^2 dP\right)^{1/2}$. We know from the Riez-Fisher Theorem that $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ is a Banach space, viz., a complete vector space. Hence, $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ is a Hilbert space.

Theorem 7.3. *Let $\{X_n\}_{n=1,2,\ldots}$ and $\{Y_n\}_{n=1,2,\ldots}$ be sequences in a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. Let $X_n \to X$ in that $\|X_n - X\| \to 0$ as $n \to \infty$ and $Y_n \to Y$. Then, $\langle X_n, Y_n\rangle \to \langle X, Y\rangle$.*

Proof. By the Cauchy-Schwarz inequality (Theorem 7.1), $|\langle X, Y\rangle| \leq \|X\|\|Y\|$. Therefore,

$$\begin{aligned}
|\langle X, Y\rangle - \langle X_n, Y_n\rangle| &= |\langle X, Y_n\rangle - \langle X_n, Y_n\rangle + \langle X, Y\rangle - \langle X, Y_n\rangle - \langle X_n, Y\rangle + \langle X_n, Y_n\rangle \\
&\quad + \langle X_n, Y\rangle - \langle X_n, Y_n\rangle| \\
&= |\langle X - X_n, Y_n\rangle + \langle X - X_n, Y - Y_n\rangle + \langle X_n, Y - Y_n\rangle| \\
&\leq |\langle X - X_n, Y_n\rangle| + |\langle X - X_n, Y - Y_n\rangle| + |\langle X_n, Y - Y_n\rangle| \\
&\leq \|X - X_n\|\|Y_n\| + \|X - X_n\|\|Y - Y_n\| + \|X_n\|\|Y - Y_n\|.
\end{aligned}$$

By convergence, $\|X - X_n\|, \|Y - Y_n\| \to 0$ and since $\|X_n\|, \|Y_n\| < \infty$ for all n, $|\langle X, Y\rangle - \langle X_n, Y_n\rangle| \to 0$, as $n \to \infty$. ∎

Definition 7.2. *Let S be a closed subset of a Hilbert space $\mathcal{H}$. The distance from $Y \in \mathcal{H}$ to S is denoted by*

$$d(Y, S) = \inf\{\|Y - X\| : X \in S\}.$$

If $Y \in S$, $d(Y,S) = 0$.

Theorem 7.4. *(Projection Theorem): Let S be a closed subspace of a Hilbert space $\mathcal{H}$ and $Y \in \mathcal{H}$. There exists a unique $X \in S$ such that $\|Y - X\| := \inf\{\|Y - X'\| : X' \in S\}$. Furthermore, $\langle Y - X, s\rangle = 0$, for all $s \in S$.*

Proof. First, consider existence of X. If $Y \in S$, put $X = Y$. If $Y \notin S$, we would like to obtain $X \in S$ such that $\|Y - X\| = \inf_{X' \in S}\{\|Y - X'\|\} = \delta > 0$.

Let $\{X_i\}_{i\in\mathbb{N}} \in S$ such that $\|X_i - Y\| \to \delta$. Now, if X_i and Y are in a Hilbert space, we have by the Parallelogram Law

$$\|(X_j - Y) + (Y - X_i)\|^2 + \|(X_j - Y) - (Y - X_i)\|^2 = 2\|X_j - Y\|^2 + 2\|Y - X_i\|^2$$

and

$$\|X_j - X_i\|^2 = 2\|X_j - Y\|^2 + 2\|Y - X_i\|^2 - 4\|Y - \frac{X_i + X_j}{2}\|^2.$$

For all i, j the vector $\frac{X_i+X_j}{2} \in S$ (since S is a subspace). Therefore, by definition of δ, $\|Y - \frac{X_i+X_j}{2}\| \geq \delta$ and we obtain $\|X_j - X_i\|^2 \leq 2\|X_j - Y\|^2 + 2\|Y - X_i\|^2 - 4\delta^2$. Since $\|X_i - Y\|^2 \to \delta^2$ by continuity of inner product (Theorem 7.3), $\|X_j - X_i\|^2 \to 0$ as $i, j \to \infty$. Hence, $\{X_i\}$ is a Cauchy sequence. Since S is closed, $\{X_i\}$ converges to $\tilde{X} \in S$. Furthermore, $\delta \leq \|Y - \tilde{X}\| \leq \|Y - X_i\| + \|X_i - \tilde{X}\| \leq \delta$. Hence, $\tilde{X} = X$ which we wanted to show existed.

Now, consider the proof of $\langle Y - X, s\rangle = 0$ for all $s \in S$. Suppose there exists $s \in S$ such that $\langle Y - X, s\rangle \neq 0$. Without loss of generality assume that $\|s\| = 1$ and that $\langle Y - X, s\rangle = \delta \neq 0$ and define $s_1 \in S$ such that $s_1 = X + \delta s$. Then,

$$\begin{aligned}
\|Y - s_1\|^2 &= \|Y - X - \delta s\|^2 \text{by definition of } s_1 \\
&= \|Y - X\|^2 - \langle Y - X, \delta s\rangle - \langle \delta s, Y - X\rangle + \delta^2\|s\|^2 \\
&= \|Y - X\|^2 - \delta^2 - \delta^2 + \delta^2 \\
&= \|Y - X\|^2 - \delta^2 < \|Y - X\|^2
\end{aligned}$$

Hence, if $\langle Y-X, s\rangle \neq 0$, then X is not the minimizing element of S and it must be that for all $s \in S$, $\langle Y-X, s\rangle = 0$.

Lastly, let's prove uniqueness. For all $s \in S$, the theorem of Pythagoras says that $\|Y-s\|^2 = \|Y-X+X-s\|^2 = \|Y-X\|^2 + \|X-s\|^2$. (Note that $\langle Y-X, X-s\rangle = 0$ due to the fact that $\langle Y-X, s\rangle = 0$, $\forall s \in S$). Hence, $\|Y-s\| > \|Y-X\|$ for $s \neq X$. ■

As a matter of terminology, we call any two elements X and Y of a Hilbert space orthogonal if $\langle X, Y\rangle = 0$.

7.2 Conditional expectation for random variables in $\mathcal{L}^2(\Omega, \mathcal{F}, P)$

Now consider the Hilbert space $\mathcal{L}^2$ composed of all random variables defined on $(\Omega, \mathcal{F}, P)$ and for precision denote this space by $\mathcal{L}^2(\Omega, \mathcal{F}, P)$. Let X be a random vector taking values in $\mathbb{R}^n$ defined in the same probability space with $\sigma(X) \subseteq \mathcal{F}$. Then, $\mathcal{L}^2(\Omega, \sigma(X), P) \subseteq \mathcal{L}^2(\Omega, \mathcal{F}, P)$ is a Hilbert space with the same inner product. Furthermore, $\mathcal{L}^2(\Omega, \sigma(X), P)$ is a closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, P)$. We now define conditional expectation.

Definition 7.3. *Let $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$. The conditional expectation of Y given X is the unique element $\hat{Y} \in \mathcal{L}^2(\Omega, \sigma(X), P)$ such that*

$$E((Y-\hat{Y})s) = 0, \text{ for all } s \in \mathcal{L}^2(\Omega, \sigma(X), P).$$

We write $\hat{Y} = E(Y|X)$ or $\hat{Y} = E(Y|\sigma(X))$.

Recall that if $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{B}^n)$ is a random vector, then $X^{-1}(\mathcal{B}^n) \subseteq \mathcal{F}$ is a σ-algebra and we wrote $X^{-1}(\mathcal{B}^n) = \sigma(X)$, the σ-algebra generated by X. Consider a random variable $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$. It is legitimate to ask when Y is measurable (a random variable) with respect to $\sigma(X)$.[2] The following theorem provides a useful characterization.

[2] More generally, for $\mathcal{G} \subset \mathcal{F}$ a σ-algebra, we say that X is $\mathcal{G}$-measurable if for all $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{G}$. There may be many of these $\mathcal{G}$'s. The intersection of all of them, i.e. $\sigma(X) := \cap_{i \in I} \mathcal{G}_i$ is called the σ-algebra generated by X.

Theorem 7.5. *Let $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{B}^n)$ be a random vector and $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ be a random variable. Y is $\sigma(X)$-measurable if, and only if, there exists $f : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$ such that $Y = f(X)$ and f is $\mathcal{B}^n$-measurable.*

Proof. ($\Longleftarrow$) We want to show that for every $B \in \mathcal{B}$ we have $Y^{-1}(B) \in \sigma(X)$. But $Y^{-1}(B) = X^{-1}(f^{-1}(B))$ and by measurability of f, $f^{-1}(B) \in \mathcal{B}^n$ and since X is a random vector $X^{-1}(f^{-1}(B)) \in \sigma(X)$. Thus, Y is $\sigma(X)$-measurable.

($\Longrightarrow$) Suppose $Y^{-1}(B) \in \sigma(X)$ for all $B \in \mathcal{B}$. First, assume that Y is simple. Then, for $k \in \mathbb{N}$ we have $Y = \sum_{i=1}^{k} a_i I_{A_i}$ for a_i distinct and A_i pairwise-disjoint. In this case, $Y^{-1}(\{a_i\}) = A_i$ and by assumption $A_i \in \sigma(X)$. Hence there exists $B_i \in \mathcal{B}^n$ such that $X^{-1}(B_i) = A_i$ (definition of $\sigma(X)$). Let $f(x) = \sum_{i=1}^{k} a_i I_{B_i}(x)$, then $Y = f(X)$, f $\mathcal{B}^n$-measurable. Thus, the implication is proved for every Y simple that is $\sigma(X)$-measurable.

If $Y : (\Omega, \mathcal{F}, P) \to [0, \infty)$ then, by Theorem 3.3, there exist $Y_n(\omega)$ simple such that

$$Y(\omega) = \lim_{n\to\infty} Y_n(\omega),\ 0 \le Y_n(\omega) \le Y_{n+1}(\omega).$$

Each Y_n is $\sigma(X)$-measurable and $Y_n = f_n(X)$ from the first part of the proof. Now, set $f(x) = \limsup_{n\to\infty} f_n(x)$ and note $Y = \lim_{n\to\infty} Y_n = \lim_{n\to\infty} f_n(X)$.

Given that $(\limsup_{n\to\infty} f_n)(X) = \limsup_{n\to\infty} f_n(X)$, by Theorem 1.20, $f(x)$ is $\mathcal{B}^n$-measurable. For general Y, write $Y = Y^+ - Y^-$ which reduces to the preceding case. ∎

Remark 7.1. *1. An equivalent way to think of Definition 7.3 using the previous theorem is to write*

$$E(Y|X) = \underset{s\in\mathcal{L}^2(\Omega,\sigma(X),P)}{arg\inf} \|Y - s\| = arg\inf_{f\in F} \|Y - f(X)\|.$$

where F is the set of Borel measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$.

2. Since $\hat{Y} = E(Y|X)$ is $\sigma(X)$-measurable, by Theorem 7.5, there exists $f : \mathbb{R}^n \to \mathbb{R}$ which is Borel measurable such that $E(Y|X) = f(X)$ and f is unique. Hence, we *can write* $E[(Y - f(X))g(X)] = 0$*, for all* $g : \mathbb{R}^n \to \mathbb{R}$ *Borel measurable such that* $\int g^2 dP < \infty$.

We can free the concept of conditional expectation from a particular set of random variables (or element) that produces $\sigma(X)$ and speak more generally of conditioning on a σ-algebra $\mathcal{G} \subset \mathcal{F}$, that is a sub-$\sigma$-algebra of $\mathcal{F}$.

Definition 7.4. $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ *be a random variable with* $\int Y^2 dP < \infty$*. Let* $\mathcal{G}$ *be a sub-*σ*-algebra of* $\mathcal{F}$*. Then* $E(Y|\mathcal{G})$ *is the unique* $\hat{Y} \in \mathcal{L}^2(\Omega, \mathcal{G}, P)$ *such that*

$$E((Y - \hat{Y})s) = E([Y - E(Y|\mathcal{G})]s) = 0,$$

for all measurable $s \in \mathcal{L}^2(\Omega, \mathcal{G}, P)$.

Remark 7.2. *1. The definition gives* $E(Ys) = E(sE(Y|\mathcal{G}))$.

2. *Since* $s = 1 \in \mathcal{L}^2(\Omega, \mathcal{G}, P)$, $E(Y) = E(E(Y|\mathcal{G}))$.

3. *If* $U, V \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$*, then* $E(U + \alpha V|\mathcal{G})$ *satisfies* $E((U + \alpha V)s) = E(E(U + \alpha V|\mathcal{G})s)$*. But,*

$$\begin{aligned} E((U + \alpha V)s) &= E(Us) + \alpha E(Vs) \\ &= E(E(U|\mathcal{G})s) + \alpha E(E(V|\mathcal{G})s) \\ &= E([E(U|\mathcal{G}) + \alpha E(V|\mathcal{G})]s). \end{aligned}$$

Hence, $E(U + \alpha V|\mathcal{G}) = E(U|\mathcal{G}) + \alpha E(V|\mathcal{G})$*. That is* $E(\cdot|\mathcal{G})$ *is a linear function.*

Theorem 7.6. *Assume that* $Z := \begin{pmatrix} Y \\ X \end{pmatrix}$ *is a random vector defined on* $(\Omega, \mathcal{F}, P)$ *taking values in* $\mathbb{R}^2$ *and having density* f.

1. Y *and* X *have densities on* $(\mathbb{R}, \mathcal{B})$ *given by* $f_Y(y) = \int_{\mathbb{R}} f(y, x) d\lambda(x)$ *and* $f_X(x) = \int_{\mathbb{R}} f(y, x) d\lambda(y)$.

2. *For every* $x \in \mathbb{R}$ *such that* $f_X(x) \neq 0$ *we have that* $f_{Y|X=x}(y) = \frac{f(y,x)}{f_X(x)}$ *is a density on* $\mathbb{R}$.

3. $E(Y|X) = h(X)$ *where* $h(x) = \int_{\mathbb{R}} y f_{Y|X=x}(y) d\lambda(y)$.

Proof. 1. Let $E \in \mathcal{B}$. Then,

$$\begin{aligned} P(Y \in E) = P(Z \in E \times \mathbb{R}) &= \int_{E \times \mathbb{R}} f(y,x) d\lambda^2(y,x) \\ &= \int_E \int_{\mathbb{R}} f(y,x) d\lambda(y) d\lambda(x) = \int_E f_Y(y) d\lambda(y) \end{aligned}$$

with $f_Y(y) = \int_{\mathbb{R}} f(y,x) d\lambda(x)$. Therefore, $P(Y \in E) = \int_{\mathbb{R}} I_E f_Y(y) d\lambda(y)$ and f_Y is a density for Y.

2. $\int_{\mathbb{R}} f_{Y|X=x}(y) d\lambda(y) = \int_{\mathbb{R}} \frac{f(y,x)}{f_X(x)} d\lambda(y) = 1$.

3. Let $h(x) = \int_{\mathbb{R}} y f_{Y|X=x}(y) d\lambda(y)$ and consider any bounded Borel measurable function $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$. Then,

$$\begin{aligned} E(h(X)g(X)) &= \int_{\mathbb{R}} h(x) g(x) f_X(x) d\lambda(x) = \int_{\mathbb{R}} \int_{\mathbb{R}} y f_{Y|X=x}(y) d\lambda(y) g(x) f_X(x) d\lambda(x) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} y \frac{f(y,x)}{f_X(x)} d\lambda(y) g(x) f_X(x) d\lambda(x) = \int_{\mathbb{R}} \int_{\mathbb{R}} y f(y,x) d\lambda(y) g(x) d\lambda(x) \\ &= E(Yg(X)) \end{aligned}$$

Consequently,

$$E(h(X)g(X)) - E(Yg(X)) = E((Y - h(X))g(X)) = 0$$

which gives $E(Y|X) = h(X)$. ■

Theorem 7.7. *Let* Y *be a random variable in* $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ *and* S *be a closed subspace of* $\mathcal{L}^2(\Omega, \mathcal{F}, P)$. *Then,*

1. *there exists a unique function* $P_S : \mathcal{L}^2(\Omega, \mathcal{F}, P) \to S$ *such that* $(\mathcal{I} - P_S) : \mathcal{L}^2(\Omega, \mathcal{F}, P) \to S^{\perp}$ *where* $S^{\perp}$ *is the orthogonal complement of* S,[3]

[3]The orthogonal complement of a subset S of an inner-product space is the set of all vectors in the space that are orthogonal to S.

2. $\|Y\|^2 = \|P_S(Y)\|^2 + \|(I - P_S)(Y)\|^2$,

3. $P_S(Y_n) \to P_S(Y)$ *if* $\|Y_n - Y\| \to 0$ *as* $n \to \infty$,

4. *if* S_1, S_2 *are closed subspaces of* $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ *such that* $S_1 \subseteq S_2 \implies P_{S_1}(P_{S_2}(Y)) = P_{S_1}(Y)$.

Proof. 1. By the Projection Theorem, for each $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ there exists a unique $\hat{Y} \in S$. Thus, we write the function $P_S(Y) = \hat{Y}$. In addition $E\{(Y - P_S(Y))s\} = 0$ for all $s \in S$. That is, $Y - P_S(Y)$ is orthogonal to the subspace S. Any $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ can be written as $Y - P_S(Y) + P_S(Y) = Y$ or $Y = (\mathcal{I} - P_S)(Y) + P_S(Y)$ where $\mathcal{I}$ is the identity operator in $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ and $\mathcal{I} - P_S$ projects Y onto the orthogonal complement of S.

2. Note that

$$\begin{aligned}\|Y\|^2 &= \|Y - P_S Y + P_S Y\|^2 \\ &= \|Y - P_S(Y)\|^2 + \|P_S(Y)\|^2 \text{ by Pythagoras' theorem} \\ &= \|(\mathcal{I} - P_S)(Y)\|^2 + \|P_S(Y)\|^2.\end{aligned}$$

3. Note that $\|P_S(Y_n) - P_S(Y)\|^2 = \|P_S(Y_n - Y)\|^2$. By the last equality in part 2.,

$$\begin{aligned}\|Y_n - Y\|^2 &= \|(\mathcal{I} - P_S)(Y_n - Y)\|^2 + \|P_S(Y_n - Y)\|^2 \\ &= \|(\mathcal{I} - P_S)(Y_n - Y)\|^2 + \|P_S(Y_n) - P_S(Y)\|^2.\end{aligned}$$

Consequently,

$$\|P_S(Y_n) - P_S(Y)\|^2 = \|Y_n - Y\|^2 - \|(\mathcal{I} - P_S)(Y_n - Y)\|^2 \le \|Y_n - Y\|^2.$$

Hence, if $\|Y_n - Y\| \to 0$ as $n \to \infty$, then $\|P_S(Y_n) - P_S(Y)\|^2 \to 0$ as $n \to \infty$.

4. $Y = P_{S_2}(Y) + (\mathcal{I} - P_{S_2})(Y)$ and $P_{S_1}(Y) = P_{S_1}(P_{S_2}(Y)) + P_{S_1}((\mathcal{I} - P_{S_2})(Y))$. In the last term, the argument of P_{S_1} is an element of the orthogonal complement of S_2. That is $<(\mathcal{I} - P_{S_2})(Y), s>= 0$ for every $s \in S_2$. But since $S_1 \subseteq S_2$, it must be that $<(\mathcal{I} - P_{S_2})(Y), s_1>= 0$ for all $s_1 \in S_1$. Thus, $(\mathcal{I} - P_{S_2})(Y) \in S_1^{\perp}$ and consequently $P_{S_1}((\mathcal{I} - P_{S_2})(Y)) = 0$. ■

In Theorem 7.7, if we take the closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ to be $\mathcal{L}^2(\Omega, \mathcal{G}, P)$ for $\mathcal{G}$ a sub σ-algebra of $\mathcal{F}$, we write $E(Y|\mathcal{G})$ for $P_S(Y)$. In particular, we have:

1. $\|Y\|^2 = \|E(Y|\mathcal{G})\|^2 + \|Y - E(Y|\mathcal{G})\|^2$,

2. $E(Y_n|\mathcal{G}) \to E(Y|\mathcal{G})$ if $Y_n \overset{\mathcal{L}^2}{\to} Y$,

3. if $\mathcal{G} \subseteq \mathcal{H}$ then $E(E(Y|\mathcal{G})|\mathcal{H}) = E(Y|\mathcal{H})$.

7.3 Conditional expectation for random variables in $\mathcal{L}(\Omega, \mathcal{F}, P)$

It is desirable to extend the concept of conditional expectation to random variables $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ such that $Y \in \mathcal{L}$. The word extend is justified, since by the Cauchy-Schwarz Inequality (or Rogers-Hölder Inequality with $p = q = 2$)

$$E(|XY|) \leq \|X\|_2 \|Y\|_2.$$

Taking $Y = 1$ almost everywhere, we have $E(|X|)^2 \leq E(X^2)$. Hence, if $E(X^2) < C$ then $E|X| < C$. Consequently, $\mathcal{L}^2 \subseteq \mathcal{L}$.

For this purpose, recall that $Y \in \mathcal{L}(\Omega, \mathcal{F}, P)$ if $Y^+ = \max\{Y(\omega), 0\}$ and $Y^- = -\min\{Y(\omega), 0\}$ are such that $E(Y^+)$, $E(Y^-) < \infty$ and, in this case, we define $E(Y) = E(Y^+) - E(Y^-)$. If $Y \geq 0$, then $Y^- = 0$ and $Y = Y^+$. We first consider $Y \in \mathcal{L}_+(\Omega, \mathcal{F}, P)$. As in Definition 3.4 we allow $Y(\omega) = \infty$. The next theorem provides the basis for extending our definition of conditional expectation to random variables in $\mathcal{L}$.

Theorem 7.8. *i) Let $Y \in \mathcal{L}_+(\Omega, \mathcal{F}, P)$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. There exists a unique element $E(Y|\mathcal{G})$ of $\mathcal{L}_+(\Omega, \mathcal{G}, P)$ such that $E([Y - E(Y|\mathcal{G})]X) = 0$ for all $X \in \mathcal{L}_+(\Omega, \mathcal{G}, P)$.*

ii) If $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ then the conditional expectation $E(Y|\mathcal{G})$ in i) is the same as $E(Y|\mathcal{G})$ in Definition 7.3 with $\sigma(X) = \mathcal{G}$.

iii) If $Y \leq Y'$ then $E(Y|\mathcal{G}) \leq E(Y'|\mathcal{G})$.

Proof. i) We first consider the existence $E(Y|\mathcal{G})$. Let $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ and $Y \geq 0$. In this case, define $E(Y|\mathcal{G})$ as in Definition 7.3. Now, for $X \in \mathcal{L}_+(\Omega, \mathcal{G}, P)$ let

$$X_n(\omega) = \min\{X(\omega), n\} = \begin{cases} X(\omega), & \text{if } X(\omega) \leq n, \\ n, & \text{if } X(\omega) > n, \end{cases}$$

and note that

$$X_n^2(\omega) = \begin{cases} X^2(\omega), & \text{if } X(\omega) \leq n \\ n^2, & \text{if } X(\omega) > n \end{cases}.$$

Hence,

$$\int_\Omega X_n^2 dP = \begin{cases} \int_\Omega X^2 dP \leq n^2 \int_\Omega dP = n^2 < \infty, & \text{if } X(\omega) \leq n \\ n^2 \int_\Omega dP = n^2 < \infty, & \text{if } X(\omega) > n \end{cases}$$

so that $X_n \in \mathcal{L}^2$.

Now, $0 \leq X_1(\omega) \leq X_2(\omega) \leq \cdots \leq X(\omega)$ and $X_n(\omega) \to X(\omega)$ almost everywhere as $n \to \infty$. Then, by Beppo-Levi's Theorem, we have that

$$E\left(\lim_{n\to\infty} YX_n\right) = E(YX) = \lim_{n\to\infty} E(YX_n) = \lim_{n\to\infty} E(E(Y|\mathcal{G})X_n).$$

The last equality follows from the fact that $EY^2 < \infty$, $EX_n^2 < \infty$ and Definition 7.3. Now, again by Beppo-Levi's Theorem, we have

$$E(YX) = \lim_{n\to\infty} E(E(Y|\mathcal{G})X_n) = E(E(Y|\mathcal{G})X), \text{ for all } X \in \mathcal{L}_+(\Omega, \mathcal{G}, P).$$

If $Y \in \mathcal{L}_+(\Omega, \mathcal{F}, P)$ then let $Y_m(\omega) = \min\{Y(\omega), m\}$ and from the argument above $Y_m \in \mathcal{L}^2$. Hence,

$$\begin{aligned} \lim_{n\to\infty} E(Y_m X_n) &= \lim_{n\to\infty} E(E(Y_m|\mathcal{G})X_n) = E(E(Y_m|\mathcal{G}) \lim_{n\to\infty} X_n) \\ &= E(E(Y_m|\mathcal{G})X). \end{aligned}$$

Now, since $Y_m \geq 0$, then $E(Y_m|\mathcal{G})$ as defined in Definition 7.3 is such that $E(Y_m|\mathcal{G}) \geq 0$. To see this, consider $Z = I_{\{E(Y_m|\mathcal{G})<0\}}$ and note that $E(Z^2) = P(E(Y_m|\mathcal{G}) < 0)$, $E(Y_m Z) =$

$E(E(Y_m|\mathcal{G})Z) = E(E(Y_m|\mathcal{G})I_{\{E(Y_m|\mathcal{G})<0\}})$. Now, since $Y_m \geq 0$ and $Z = 1$ or $Z = 0$ we have that $E(Y_m Z) \geq 0$. But the right-hand side of the last equality is less than 0 if $E(Y_m|\mathcal{G}) < 0$, so it must be that $E(Y_m|\mathcal{G}) \geq 0$ if $Y_m \geq 0$. Hence, $E(Y_m|\mathcal{G})$ is increasing with m, and by Beppo-Levi's Theorem we have

$$\lim_{m\to\infty}\lim_{n\to\infty} E(Y_m X_n) = E(YX) = \lim_{m\to\infty} E\left(E(Y_m|\mathcal{G})X\right) = E\left(\lim_{m\to\infty} E(Y_m|\mathcal{G})X\right).$$

Now, since $E(YX) = E\left(\left(\lim_{m\to\infty} E(Y_m|\mathcal{G})\right)X\right)$ or $E\left(\left(Y - \lim_{m\to\infty} E(Y_m|\mathcal{G})\right)X\right) = 0$ for all $X \in \mathcal{L}_+(\Omega, \mathcal{G}, P)$, we define

$$E(Y|\mathcal{G}) = \lim_{m\to\infty} E(Y_m|\mathcal{G}) \tag{7.1}$$

for $Y \in \mathcal{L}^+(\Omega, \mathcal{F}, P)$.

We now consider uniqueness of $E(Y|\mathcal{G})$. Let U and V be two versions of $E(Y|\mathcal{G})$ and let $\wedge_n = \{\omega : U < V \leq n\}$. Since U and V are versions of $E(Y|\mathcal{G})$ we know that U and V are $\mathcal{G}$-measurable. Consequently, $\{\omega : U \leq n\} \in \mathcal{G}$, $\{\omega : V \leq n\} \in \mathcal{G}$ and $\wedge_n = \{\omega : U < V \leq n\} \in \mathcal{G}$.

Note that $E(YI_{\wedge_n}) = E(UI_{\wedge_n}) = E(VI_{\wedge_n})$ since $U = V = E(Y|\mathcal{G})$. Furthermore, $0 \leq UI_{\wedge_n} \leq VI_{\wedge_n} \leq n$ and if $P(\wedge_n) > 0$ ($\wedge_n \neq \emptyset$), $UI_{\wedge_n} < VI_{\wedge_n}$ which implies that $E(UI_{\wedge_n}) < E(VI_{\wedge_n})$, which contradicts $E(UI_{\wedge_n}) = E(VI_{\wedge_n})$. Therefore, $P(\wedge_n) = 0$ for all n. Now, note that $\wedge_1 \subseteq \wedge_2 \subseteq \wedge_3 \subseteq \cdots \subseteq \{U < V\}$. Now $\lim_{n\to\infty} \cup_{i=1}^n \wedge_i = \{U < V\}$ and $P\left(\lim_{n\to\infty} \cup_{i=1}^n \wedge_i\right) = \lim_{n\to\infty} P\left(\cup_{i=1}^n \wedge_i\right) \leq \lim_{n\to\infty} \sum_{i=1}^n P(\wedge_i)$. Thus, $P(\{U < V\}) = 0$. Repeating the argument for $\Gamma_n = \{\omega : V < U \leq n\}$ we conclude that $P(\{V < U\}) = 0$. Hence, it must be that U and V coincide with probability 1.

ii) The proof follows from the first part of the argument in item i).

iii) If $Y \leq Y'$ then $Y_m \leq Y'_m$ for all m and $E(Y_m|\mathcal{G}) \leq E(Y'_m|\mathcal{G})$ and consequently

$$\lim_{m\to\infty} E(Y_m|\mathcal{G}) \leq \lim_{m\to\infty} E(Y'_m|\mathcal{G}) \iff E(Y|\mathcal{G}) \leq E(Y'|\mathcal{G}).$$

■

We now consider conditional expectations for random variables in $\mathcal{L}(\Omega, \mathcal{F}, P)$.

Theorem 7.9. *Let $Y \in \mathcal{L}(\Omega, \mathcal{F}, P)$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. There exists a unique element $E(Y|\mathcal{G})$ in $\mathcal{L}(\Omega, \mathcal{G}, P)$ such that*

$$E\left(\left(Y - E(Y|\mathcal{G})\right)X\right) = 0, \textit{ for all bounded } \mathcal{G}\textit{-measurable } X.$$

$E(Y|\mathcal{G})$ coincides with those in Definition 7.3 and Theorem 7.8 when $Y \in \mathcal{L}^2$ and $Y \in \mathcal{L}_+$. In addition, (i) if $Y \geq 0$, then $E(Y|\mathcal{G}) \geq 0$ and (ii) $E(Y|\mathcal{G})$ is a linear in Y.

Proof. We first consider existence of the conditional expectation. Since $Y \in \mathcal{L}$, we can write $Y = Y^+ - Y^-$ and $Y^+, Y^- \in \mathcal{L}$. Now, Y^+ and Y^- are such that

$$E\Big((Y^+ - E(Y^+|\mathcal{G}))X\Big) = 0, \text{ for all } X \in \mathcal{L}_+(\Omega, \mathcal{G}, P) \text{ and}$$
$$E\Big((Y^- - E(Y^-|\mathcal{G}))X\Big) = 0, \text{ for all } X \in \mathcal{L}_+(\Omega, \mathcal{G}, P).$$

Define $E(Y|\mathcal{G}) = E(Y^+|\mathcal{G}) - E(Y^-|\mathcal{G})$ and note that for $X \in \mathcal{L}_+(\Omega, \mathcal{G}, P)$

$$\begin{aligned} E(YX) &= E((Y^+ - Y^-)X) = E(Y^+X) - E(Y^-X) \\ &= E(E(Y^+|\mathcal{G})X) - E(E(Y^-|\mathcal{G})X) \text{ by Theorem 7.8} \\ &= E((E(Y^+|\mathcal{G}) - E(Y^-|\mathcal{G})))X) = E(E(Y|\mathcal{G})X). \end{aligned}$$

We now establish uniqueness of $E(Y|\mathcal{G})$. Suppose U and V are two versions of $E(Y|\mathcal{G})$ and let $\wedge = \{U < V\}$. Then, since U and V are $\mathcal{G}$-measurable, then $\wedge \in \mathcal{G}$. Therefore $I_\wedge$ is $\mathcal{G}$-measurable.

$$E(YI_\wedge) = E(E(Y|\mathcal{G})I_\wedge) = E(UI_\wedge) = E(VI_\wedge).$$

But, if $P(\wedge) > 0$, then $E(UI_\wedge) < E(VI_\wedge)$, a contradiction. Thus, $P(\wedge) = 0$. A similar reverse argument gives $P(V < U) = 0$.

Now, for any X that is bounded and $\mathcal{G}$-measurable consider

$$E(YX) = E(Y(X^+ - X^-)) = E(YX^+) - E(YX^-)$$
$$= E(X^+E(Y|\mathcal{G})) - E(X^-E(Y|\mathcal{G}))$$

using the definition of conditional expectation in this proof.

$$= E((X^+ - X^-)E(Y|\mathcal{G})) = E(XE(Y|\mathcal{G})).$$

The proofs of items (i) and (ii) are left as exercises. ■

Remark 7.3. *Note that if X and Y are independent random variables defined on the same probability space, then by Theorem 4.7, if f is a bounded measurable function $E(Yf(X)) = E(Y)E(f(X))$. Now, $E(Yf(X)) = E(E(Y|\sigma(X))f(X))$ and consequently*

$$E(Y)E(f(X)) = E(E(Y|\sigma(X))f(X)),$$

taking $f(X) = 1$ gives $E(Y) = E(Y|\sigma(X))$.

Lebesgue's monotone and dominated convergence theorems hold for conditional expectations.

Theorem 7.10. *$Y_n(\omega) : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$.*

a) If $Y_n \geq 0$, $Y_1 \leq Y_2 \leq Y_3 \leq \ldots$ with $Y_n \stackrel{as}{\to} Y$ as $n \to \infty$, then $\lim_{n\to\infty} E(Y_n|\mathcal{G}) = E(Y|\mathcal{G})$ a.s.

b) If $Y_n \stackrel{as}{\to} Y$ and $|Y_n| \leq Z$ for some $Z \in \mathcal{L}(\Omega, \mathcal{F}, P)$, then $\lim_{n\to\infty} E(Y_n|\mathcal{G}) = E(Y|\mathcal{G})$ a.s.

Proof. Left as an exercise. ■

We now give an example where conditional expectation is taken to belong to a specific class of measurable functions.

Example 7.3. *Let $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$ and let X be a random vector defined on the same probability space. Assume that for every component of X_k, for $k = 1, \cdots, K$ of X we have $X_k \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$. Now, consider the following class of functions*

$$F = \{f : f(x) = \sum_{k=1}^{K} a_k x_k \text{ where } f \text{ is } \sigma(X)\text{-measurable and } a_k \in \mathbb{R}\}.$$

Using Definition 7.3 or item 1 in Remark 30

$$E(Y|X) = \underset{a_1,\dots,a_K}{argmin} \int \left(Y - \sum_{k=1}^{k} a_k X_k \right)^2 dP = \underset{a_1,\dots,a_K}{argmin}\, O(a_1, \dots, a_K).$$

Now,

$$\begin{aligned}
O(a_1, \dots, a_K) &= \int (Y^2 - 2Y \sum_{k=1}^{K} a_k X_k + (\sum_{i=1}^{K} a_k X_k)^2) dP \\
&= \int Y^2 dP - 2 \sum_{k=1}^{K} a_k \int X_k Y dP + \sum_{k=1}^{K} a_k^2 \int X_k^2 dP \\
&+ \sum_{k=1}^{K} \sum_{k \neq l} a_k a_l \int X_k X_l dP \\
&= \sigma^2 - 2 \sum_{k=1}^{K} a_k E(X_k Y) + \sum_{k=1}^{K} a_k^2 \int X_k^2 dP + \sum_{k=1}^{K} \sum_{jk \neq l} a_k a_l E(X_k X_l).
\end{aligned}$$

Now, taking derivatives with respect to a_k we have $\frac{\partial}{\partial a_k} O(a_1, \dots, a_K) = -2E(X_k Y) + 2a_k E(X_k^2) + 2\sum_{k \neq l} a_l E(X_k X_l)$ for $k = 1, \dots, K$. Alternatively, using matrices

$$\begin{aligned}
\frac{\partial}{\partial a} O(a_1, \dots, a_K) &= -2 \begin{bmatrix} E(X_1 Y) \\ \vdots \\ E(X_K Y) \end{bmatrix} + 2 \begin{bmatrix} E(X_1^2) & E(X_1 X_2) & \dots & E(X_1 X_K) \\ E(X_2 X_1) & E(X_2^2) & \dots & E(X_2 X_K) \\ \vdots & & \vdots & \\ E(X_K X_1) & E(X_K X_2) & \dots & E(X_K^2) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix} \\
&= -2b + 2Aa
\end{aligned}$$

Choosing $a := \hat{a}$ such that $\frac{\partial}{\partial a} O(\hat{a}_1, \dots, \hat{a}_K) = 0$ we have $\hat{a} = A^{-1} b$ if A is invertible. Invertibility of A follows positive definiteness of A, which also assures that $\hat{f}(x) = \sum_{k=1}^{K} \hat{a}_k x_k$ corresponds to a minimum.